

Connecting protein and mRNA burst distributions for stochastic models of gene expression

Vlad Elgart[‡], Tao Jia, Andrew T. Fenley and Rahul Kulkarni

Department of Physics, Virginia Tech, Blacksburg, VA 24061

E-mail: elgart@vt.edu, kulkarni@vt.edu

Abstract. The intrinsic stochasticity of gene expression can lead to large variability in protein levels for genetically identical cells. Such variability in protein levels can arise from infrequent synthesis of mRNAs which in turn give rise to bursts of protein expression. Protein expression occurring in bursts has indeed been observed experimentally and recent studies have also found evidence for transcriptional bursting, i.e. production of mRNAs in bursts. Given that there are distinct experimental techniques for quantifying the noise at different stages of gene expression, it is of interest to derive analytical results connecting experimental observations at different levels. In this work, we consider stochastic models of gene expression for which mRNA and protein production occurs in independent bursts. For such models, we derive analytical expressions connecting protein and mRNA burst distributions which show how the functional form of the mRNA burst distribution can be inferred from the protein burst distribution. Additionally, if gene expression is repressed such that observed protein bursts arise only from single mRNAs, we show how observations of protein burst distributions (repressed and unrepressed) can be used to completely determine the mRNA burst distribution. Assuming independent contributions from individual bursts, we derive analytical expressions connecting means and variances for burst and steady-state protein distributions. Finally, we validate our general analytical results by considering a specific reaction scheme involving regulation of protein bursts by small RNAs. For a range of parameters, we derive analytical expressions for regulated protein distributions that are validated using stochastic simulations. The analytical results obtained in this work can thus serve as useful inputs for a broad range of studies focusing on stochasticity in gene expression.

Keywords stochastic, gene expression, regulation, small RNA, bursts.

PACS numbers: 87.10Mn, 87.18Tt

Submitted to: *Phys. Biol.*

[‡] Present Address: Department of Microbiology and Immunology, Baxter Lab for Stem Cell Biology, Stanford University, School of Medicine, Palo, Alto, California

1. Introduction

The intrinsic stochasticity of biochemical reactions has important consequences for the functioning of cellular processes [1, 2]. In particular, reactions corresponding to the process of gene expression often involve small numbers of molecules, and can be subject to large fluctuations. The corresponding stochasticity in gene expression has been identified as a key factor underlying the observed phenotypic variability of genetically identical cells in homogeneous environments [3]. Quantifying the effects of intrinsic noise using stochastic models of gene expression is thus an important step towards understanding cellular function and variability.

Several recent studies have focused on quantifying noise in gene expression using both single-cell assays and single-molecule techniques. Experimental observations of noise in steady-state protein distributions across a population of cells [4] were shown to be consistent with predictions from simple models based on translation from individual mRNAs [4, 5]. These models predict that each mRNA produces a burst of protein that is geometrically distributed [6]. Single-molecule studies have indeed seen protein production occurring in bursts and determined that the corresponding protein burst distribution is geometric [7, 8, 9]. At the mRNA level, single-molecule studies have demonstrated that mRNA production can also occur in transcriptional bursts [3, 7, 10, 11, 12, 13]. The presence or absence of transcriptional bursting indicates different sources of noise in gene expression and several studies are currently engaged in probing gene expression at multiple stages to elucidate the underlying sources of variability [13, 14, 15].

Given different experimental techniques for probing stochasticity in gene expression using measurements at different stages (specifically steady-state and burst distributions for proteins and mRNAs) [16, 17], it is of interest to derive analytical results connecting observables at different levels. These results can be used to infer information at one level using experiments at a different level. For example, in previous work [18] it was shown that experimental determination of the protein burst distribution and frequency can be used to determine the steady-state protein distribution. In this context, we note that most previous models have focused on reaction schemes which correspond to a geometric burst distribution for proteins produced from a single mRNA [18, 19]. However, more general reaction schemes for protein production from mRNAs can lead to deviations from geometric burst distributions for single mRNA bursts [20]. It would thus be desirable to derive analytical formulae connecting burst and steady-state protein distributions for arbitrary protein burst distributions. Finally, we note that such analytical results can be used to check for consistency between the experimental results from probing different levels of gene expression. In particular, any observed inconsistencies could signal that some model assumptions are invalid, potentially leading to new insights about the mechanisms of gene expression.

In this work, we analyze a class of burst models for protein production from mRNAs and derive analytical results connecting observable distributions at different stages of

gene expression. In particular, we show how the functional form of the mRNA burst distribution can be determined using the observed protein burst distribution. If mRNA transcription can be repressed such that observed protein bursts arise only from single mRNAs, then the derived results show how observations of protein burst distributions (repressed and unrepressed) can be used to completely determine the mRNA burst distribution. Assuming independent bursts whose arrival can be modeled as a Poisson process, we derive expressions connecting the mean and variance of protein burst distributions to the corresponding quantities for the steady-state distribution of protein levels across a population of cells. Finally, we consider a specific example for which burst distributions can deviate from the geometric distribution: post-transcriptional regulation of bursts by small RNAs. In the limit of low burst frequency, we derive analytical expressions for the protein burst distribution which are in excellent agreement with results from stochastic simulations. The results derived in this work can thus serve as useful building blocks for future studies focusing on stochasticity in gene expression.

2. Tools, Notations, and Definitions

A starting point of our analysis is the Master equation [21]

$$\partial_t P(\vec{n}; t) = \hat{H}(\vec{n}) P(\vec{n}; t), \quad (1)$$

where $\vec{n} = \{n_X\}$ is a state vector describing the abundance of each species X in the system. Here $P(\vec{n}; t)$ is the probability to find the system with the state vector \vec{n} after time t has elapsed. Equation (1) is supplemented by the initial conditions, namely the initial distribution $P_0(\vec{n})$ at time $t = 0$.

The generating function of the probability distribution Eq. (1) is defined by

$$G(\vec{x}; t) \equiv \sum_{n_i=0}^{\infty} P(\vec{n}; t) \prod_{i=\{X\}} x_i^{n_i}, \quad (2)$$

where $\vec{x} = \{x_i\}$ is a real vector dual to the state vector \vec{n} . The generating function, in turn, satisfies the corresponding evolution equation and initial condition

$$\partial_t G(\vec{x}; t) = \hat{\mathcal{H}}(\vec{x}) G(\vec{x}; t), \quad (3)$$

$$G(\vec{x}; 0) = G_0(\vec{x}) \equiv \sum_{n_i=0}^{\infty} P_0(\vec{n}) \prod_i x_i^{n_i}. \quad (4)$$

Let us first consider the simplest gene expression reaction scheme. The minimal model [6] of gene expression is given by the diagram on Fig.1. The corresponding reaction scheme is



where D is DNA, M is mRNA, and P is protein. Both mRNAs and proteins are synthesized at the constant rates k_m and k_p respectively, and their degradation (decay)

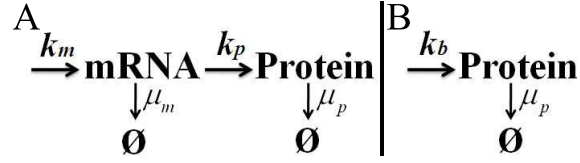


Figure 1. Reaction scheme for the minimal model of gene expression. *A.* mRNA transcripts are created with reaction rate k_m and degraded with reaction rate μ_m . Protein is translated from mRNA with a reaction rate k_p and degraded with a reaction rate μ_p . *B.* Evolution of protein distributions for the same scheme can be analyzed in terms of arrivals of bursts of proteins (valid when $\mu_p \ll \mu_m$) [18].

rates are μ_m and μ_p . Correspondingly, the evolution operator in the equation for the generating function is given by

$$\begin{aligned}
 \hat{\mathcal{H}} = & k_m(x_m - 1) + \mu_m(1 - x_m) \frac{\partial}{\partial x_m} + \\
 & k_p(x_p - 1)x_m \frac{\partial}{\partial x_m} + \mu_p(1 - x_p) \frac{\partial}{\partial x_p}.
 \end{aligned} \tag{6}$$

Note that generating function representation is particularly useful since it converts the infinite set of equations for various integer values of \vec{n} in Eq.(1) into a single partial differential equation. Moreover, a calculation of any observable quantities, such as moments of distribution, is equivalent to evaluation of derivatives of the generating function at point $\vec{x} = \{1, 1, \dots, 1\}$.

There are several parameters that describe the dynamics of the gene expression models analyzed in this work. The following rules serve as general guides for the notation used:

- Indices $X = m, p, s$ stand for mRNA, protein, and sRNA species correspondingly.
- Lower case letters are used to describe burst variable, e.g., p_m denotes the probability distribution of mRNA burst size and g_m is its generating function.
- Capital letters are used to describe steady-state variable, e.g., P_p denotes the protein steady-state distribution and G_p is its generating function.

Finally, we define some distributions that arise when considering bursts of gene expression. The geometric distribution is given by

$$\tilde{\rho}(n) = (1 - u)^n u, \quad n \geq 0 \tag{7}$$

with the corresponding generating function

$$\tilde{G}(x) = \frac{u}{1 - (1 - u)x}. \tag{8}$$

and mean given by $(1 - u)/u$. It is also convenient to define the *conditional* geometric distribution

$$\begin{aligned}
 \rho(0) &= 0, \\
 \rho(n) &= (1 - u)^{n-1} u, \quad n \geq 1
 \end{aligned} \tag{9}$$

with the corresponding generating function

$$G(x) = \frac{u x}{1 - (1 - u) x}. \quad (10)$$

The conditional geometric distribution is encountered [19] when considering the distribution of mRNAs that give rise to a protein burst. Since the observation of a burst of proteins necessarily implies the presence of at least 1 mRNA, the distribution is conditioned accordingly. We note that the mean of the conditional distribution is given by $1/u$. In the limit $u \rightarrow 1$, the distribution Eq. (9) describes a mRNA burst with exactly 1 mRNA produced per burst, which is the case when mRNA arrival corresponds to a Poisson process.

3. Bursts and modeling framework

Recent experiments have determined the variation of noise in protein expression as a function of mean protein abundance for several genes [22, 23]. The observed scaling relationship is consistent with different underlying models (see Figure 2). In one case, the transcription rate k_m is constant corresponding to a Poisson process driving mRNA synthesis. Another possible scenario corresponds to a Telegraph process [2, 10, 24, 25, 26] driving the creation of mRNAs. In this case, the promoter driving gene expression switches between active and inactive states. When the promoter is in the active state, multiple number of mRNAs can be transcribed. While both models are consistent with the experimental data, the observed scaling indicates that protein production occurs in infrequent bursts for many genes.

Based on observations relating to bursts, an analytical approach [18] was introduced to derive expressions for steady-state protein distributions from protein burst distributions. Specifically, it is assumed that (i) protein degradation rate is much smaller than mRNA degradation rate ($\mu_p \ll \mu_m$), (ii) protein levels vary due to *independent* bursts of protein expression in combination with changes due to protein degradation and (iii) the arrival of bursts can be modeled as a Poisson process. The above approach then reduces the problem of characterizing protein steady-state distributions into two parts: (i) first obtain the protein burst distribution for a single burst and (ii) using this burst distribution as input, derive and analyze the corresponding Master equation (see Fig. 1B) for proteins alone [18, 27]. A mathematical justification of this procedure of deriving a Master equation for proteins alone, given the assumptions stated above, has been provided recently [27].

In the following sections, we will consider stochastic models of gene expression consistent with the assumptions stated above. Specifically, we consider models for which mRNA and protein production occurs in independent bursts such that the arrival of bursts corresponds to a Poisson process. As noted above, even for a Poisson process, there are parameter constraints that must be satisfied ($\mu_p \ll \mu_m$) for the burst approximation to be valid. While previous work has largely focused on reaction schemes that give rise to a geometric burst of proteins from a single mRNA, we will consider

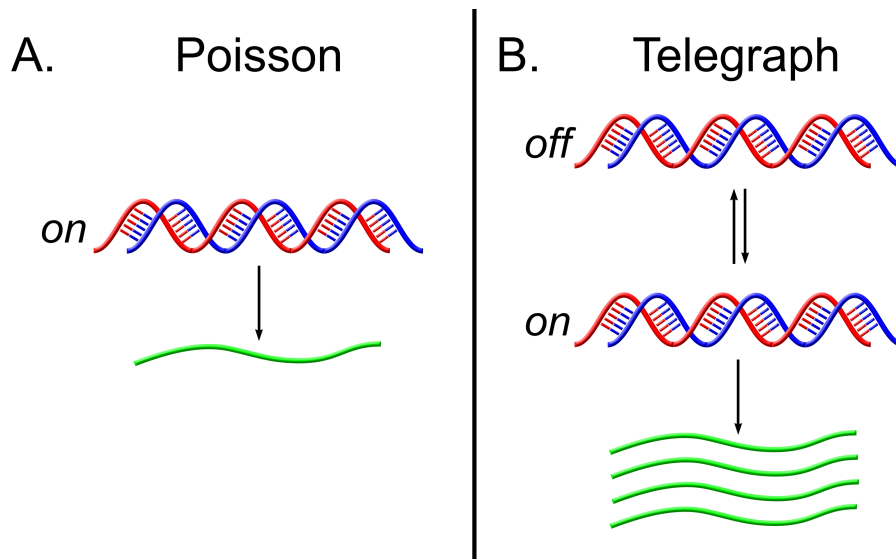


Figure 2. (color online) Schematic representation of the Poisson and Telegraph mRNA transcription processes. *A.* The Poisson process of transcription. The DNA is always in the “on” state resulting in a constant production rate of mRNA transcript (green lines). *B.* The Telegraph process of transcription. The DNA exchanges between two states, “off” and “on”. The “off” state corresponds to inactive DNA in which no transcripts are produced, while the “on” state corresponds to DNA capable of producing a burst of mRNA transcripts before it reverts back to the “off” state.

the case for which the protein burst distribution can be an arbitrary function. For such models, we wish to derive analytical expressions which connect observations at different stages of gene expression (see Figure 3). The following section first considers how observations of protein burst distributions can inform us about the underlying mRNA burst distributions.

4. From protein to mRNA burst distribution

We first consider the minimal scheme (Eq. (5)) of protein production from mRNAs. For this scheme, the following equation relating the mRNA and protein burst distributions can be derived (see Appendix):

$$g_m(x) = g_p \left[1 - \frac{(1-x)}{(k_p/\mu_m)x} \right], \quad (11)$$

The functions g_m and g_p in the equation Eq. (11) are correspondingly the mRNA and protein burst generating functions. The dynamical version (for time dependent distributions) of Eq. (11) can also be found in the Appendix. Note that, consistent with the assumption $\mu_m \gg \mu_p$, we ignore protein degradation during a single burst, i.e. the above equation considers only the proteins synthesized during the burst.

The result Eq. (11) is useful because it allows us to infer the functional form of the mRNA burst distribution from observations of protein burst distributions. Consider the

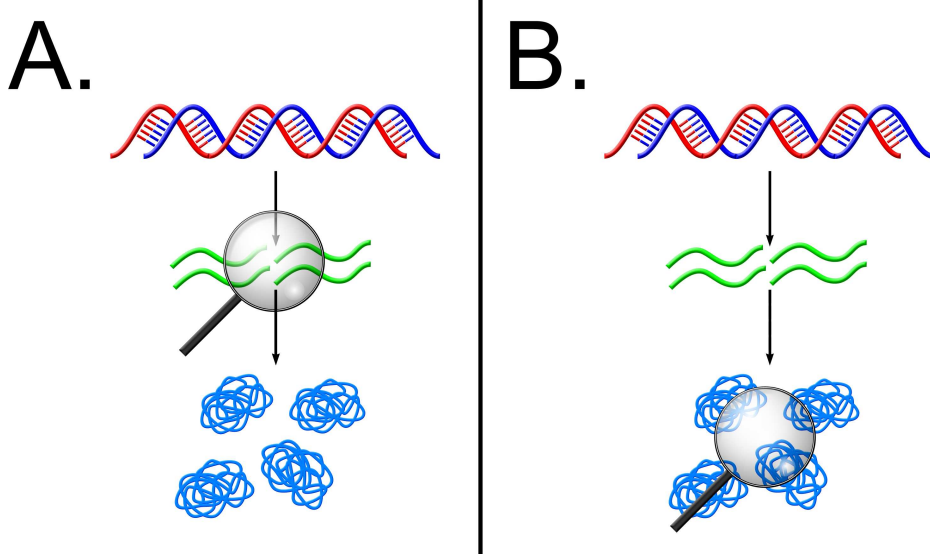


Figure 3. Different approaches for probing bursts of gene expression. *A.* Measuring the mRNA burst distributions directly. *B.* Measuring the protein distributions resulting from the mRNA burst distributions. The derived results provide means of connecting these two measurements at different stages of gene expression for the case of arbitrary protein burst distributions arising from a single sRNA.

case that the observed protein burst distribution (e.g. as reported in [8]) is a geometric distribution (Eq. (7)) with parameter $u = u_p$. Then, using the expression Eq. (11), we obtain that the mRNA burst distribution has to be a conditional geometric distribution, Eq. (9), with parameter

$$u_m = \frac{k_p}{\mu_m} \frac{u_p}{1 - u_p}. \quad (12)$$

In other words, the mRNA burst distribution is given by

$$P_m(n) = \rho(u_m), \quad n \geq 1 \quad (13)$$

While the functional form of the mRNA burst distribution is thus determined, we note that the precise distribution is not known since the parameter $(\frac{k_p}{\mu_m})$ is not known. The upper bound for $\frac{k_p}{\mu_m}$ can be derived from the condition $u_m = 1$

$$\left(\frac{k_p}{\mu_m} \right)_{\max} = \frac{1 - u_p}{u_p}, \quad (14)$$

which corresponds to the Poisson scenario, i.e. the observed burst distribution is produced from a single mRNA. On the other hand, we can have $(\frac{k_p}{\mu_m}) < (\frac{k_p}{\mu_m})_{\max}$, which implies $u_m < 1$ and thereby that the mean number of mRNAs in the burst ($\frac{1}{u_m}$) is greater than 1. This set of parameters would be consistent with a Telegraph process driving mRNA creation since it produces a geometric mRNA burst distribution (with $u_m < 1$) and also gives rise to a geometric protein burst distribution [19].

It has been noted in previous work [10, 22, 19] that the Poisson and Telegraph processes cannot be distinguished by experimental observations on proteins alone, since

both both Poisson and Telegraph processes give rise to a geometric burst distribution for proteins. However, previous work did not preclude other possible mRNA burst distributions that can result in a geometric protein burst distribution. The preceding analysis demonstrates that, if the observed protein burst distribution is geometric, then the mRNA burst distribution has to be a conditional geometric distribution. Thus Eq. (13) is a mathematically necessary and sufficient condition on the mRNA burst distribution to obtain a geometric burst distribution for proteins. An important corollary is that kinetic schemes which lead to non-geometric mRNA burst distributions can be ruled out if the observed protein burst distribution is a geometric distribution.

Let us now consider general reaction schemes which can give rise to non-geometric protein burst distributions. This can occur due to interaction with a post-transcriptional regulator [20] or even otherwise, e.g. if we have switching between competing mRNA secondary structure conformations which correspondingly have different protein production and/or mRNA degradation rates. Another example is the case for which mRNA degradation is not a Poisson process but occurs in stages (termed mRNA senescence [28]); in general the corresponding protein burst distribution will not be a geometric distribution. In such cases, the preceding analysis can be generalized as follows. Let us denote by $\phi(x)$ the generating function of protein bursts obtained from a *single* mRNA. The number of proteins produced in a single burst can be expressed as the sum of a random number (N) of random variables, each of which is drawn from the probability distribution corresponding to $\phi(x)$. The random variable N corresponds to the number of mRNAs in the burst with generating function $g_m(x)$. Correspondingly, the generating function of the protein burst distribution is given by

$$g_p(x) = g_m[\phi(x)], \quad (15)$$

Inversion of Eq. (15) yields the mRNA burst distribution

$$g_m(z) = g_p[\phi^{-1}(z)], \quad (16)$$

Note that Eq. (11) is a special case of Eq. (16). For the minimal scheme of gene expression (Fig. 1), the burst distribution from a single mRNA is a geometric distribution with mean $\frac{k_p}{\mu_m}$ [6]. Correspondingly, the generating function is given by $\phi(x) = \frac{u}{1-(1-u)x}$, with $u = \frac{\mu_m}{k_p + \mu_m}$. Inversion of $\phi(x)$ in combination with Eq. (16) gives Eq. (11).

The significance of the above equations is that once $\phi(x)$ is determined, the mRNA burst distribution can be inferred from the observed protein burst distribution (and vice-versa). Recent experiments [8] have shown that repressors can be used to regulate gene expression such that each observed burst corresponds to proteins produced from a single mRNA. Such experiments can be used to determine the single mRNA burst distribution and hence $\phi(x)$. Thus, if the protein burst distributions can be observed for both scenarios, with and without the repressor, then Eq. (16) can be used to completely determine the mRNA burst distribution.

5. Connecting Burst and Steady-State Distributions

While the direct observation of protein expression bursts has been demonstrated experimentally [7, 8, 9]; in general, carrying out such experiments is challenging. Since steady-state protein distributions are less challenging to determine experimentally, it is of interest to derive results connecting burst and steady-state distributions, in particular connecting the means and variances. We note that recent work [28] has derived results connecting burst and steady-state variances for general models of gene expression, in particular for models such that the waiting-time distribution between bursts can be arbitrary, as opposed to the simple exponential distribution which corresponds to a Poisson process for burst arrival. In the following, we first focus on the case of Poisson arrivals for bursts.

As discussed in Section 3, we assume that each burst can be considered as an independent realization of the same stochastic process and that burst arrival can be modeled as a Poisson process. Let us denote by $P_b(n)$ the probability that n proteins are produced during a single burst. Correspondingly, the Master equation for the protein distribution at time t ($P(n, t)$) is [29]:

$$\begin{aligned} \partial_t P(n, t) = & \mu_p [P(n+1, t) - P(n, t)] \\ & + k_b \sum_{n'=0}^{\infty} [P_b(n') P(n-n', t) - P_b(n') P(n, t)] \end{aligned} \quad (17)$$

The parameter k_b is the constant rate of burst arrival, i.e. it is the inverse of the mean time between two sequential bursts. If each burst corresponds to proteins produced from a single mRNA, then k_b is identical to the mRNA creation rate k_m .

Let us define the generating functions:

$$G_b(x) = \sum_{n=0}^{\infty} x^n P_b(n), \quad (18)$$

$$G(x, t) = \sum_{n=0}^{\infty} x^n P(n, t) \quad (19)$$

Correspondingly, the evolution equation for the generating function is [29]

$$\partial_t G(x, t) = \mu_p (1-x) \partial_x G(x, t) + k_b [G_b(x) - 1] G(x, t), \quad (20)$$

The time dependent solution of Eq.(20) can be obtained by the method of characteristics. The steady-state limit ($G_s(x)$) is given by [29],

$$G_s(x) = \exp \left\{ \frac{k_b}{\mu_p} \int_1^x \left(\frac{G_b(y) - 1}{y - 1} \right) dy \right\}. \quad (21)$$

From Eq.(21) we can also derive useful expressions for the mean and the Fano factor (or noise strength) of the steady-state distribution in terms of the corresponding quantities for burst distribution. We obtain:

$$\bar{n}_s = \left(\frac{k_b}{\mu_p} \right) \bar{n}_b, \quad (22)$$

$$\frac{\sigma_s^2}{\bar{n}_s} = 1 + \bar{n}_b + \frac{1}{2} \left(\frac{\sigma_b^2}{\bar{n}_b} - (1 + \bar{n}_b) \right). \quad (23)$$

If the burst distribution is geometric, we have $\frac{\sigma_b^2}{\bar{n}_b} = 1 + \bar{n}_b$ and the above result reduces to previously obtained results [2, 27] in the limit $\mu_m \gg \mu_p$. For general reaction schemes, the burst distribution differs from the geometric distribution and Eq. (23) is the generalization that connects burst and steady-state distributions. It is interesting to note that a similar result was obtained in previous work [28] with different model assumptions: specifically, each mRNA was assumed to produce a geometric burst of proteins, however the number of mRNAs in the burst was assumed to be drawn from an arbitrary burst distribution.

The preceding discussion focused on the case that the burst arrival is a Poisson process, thus the waiting-time distribution between bursts is given by an exponential distribution. For a Poisson process driving mRNA production this is certainly the case. However, mRNA production has also been proposed to arise from a Telegraph process [10, 13] which, in general, does not have the feature that the waiting-time between bursts is an exponential distribution [26]. We consider the case that the DNA fluctuates between two different conformations which correspond to different production rates for the mRNAs. In particular, we consider a two-stage model [10] corresponding to two different active conformations of DNA (“on” and “off” or 1 and 2 say), with mRNA transcription rates k_1 and $k_2 (< k_1)$ respectively. Note that previous work has focused on the case $k_2 = 0$, i.e. no transcription in the “off” state. The present results generalize this model to allow for a basal level of transcription in the “off” state as well. Let us define a parameter $f = \frac{k_2}{k_1}$ which is the ratio of the two rates and takes values between 0 and 1. We denote by λ_{12} the rate of switching from conformation 1 to conformation 2, and λ_{21} is the rate for the reversed process. For this two-stage model, an analytical expression linking the burst distribution to the steady-state distribution (analogous to Eq. (21)) seems intractable. However, we can derive expressions for steady-state mean and variance (see Appendix). We obtain:

$$\bar{n}_s = \frac{\bar{k}}{\mu_p} \bar{n}_b, \quad (24)$$

$$\begin{aligned} \frac{\sigma_s^2}{\bar{n}_s} = 1 + \bar{n}_b + \frac{1}{2} \left(\frac{\sigma_b^2}{\bar{n}_b} - (1 + \bar{n}_b) \right) \\ + \left(1 + \frac{\lambda_{12} + \lambda_{21}}{\mu_p} \right)^{-1} \left(\frac{\lambda_{12}}{\lambda_{21}} \right) \left[\frac{1 - f}{1 + \frac{\lambda_{12}}{\lambda_{21}} f} \right]^2 \bar{n}_s, \end{aligned} \quad (25)$$

where we defined

$$\bar{k} = \frac{\lambda_{21} k_1 + \lambda_{12} k_2}{(\lambda_{12} + \lambda_{21})}. \quad (26)$$

Note that for the case $f = 0$, the above formula reduces to previously obtained results [2, 27], whereas for $f = 1$ we recover Eq. (23). The above result thus generalizes

previously obtained results for the case of nonzero f and for arbitrary protein burst distributions.

6. Burst Distribution for Regulation by Small RNAs

The preceding sections derived general results connecting burst and steady-state distributions for burst distributions which can deviate from a geometric distribution. We now consider a specific regulation scheme that can give rise to non-geometric protein burst distributions: regulation by small RNAs. Small RNAs are genes that are transcribed but not translated, i.e. they are non-coding RNAs. In bacteria, small RNAs have been studied extensively in recent years [30] in part due to the critical roles they play in cellular post-transcriptional regulation in response to environmental changes.

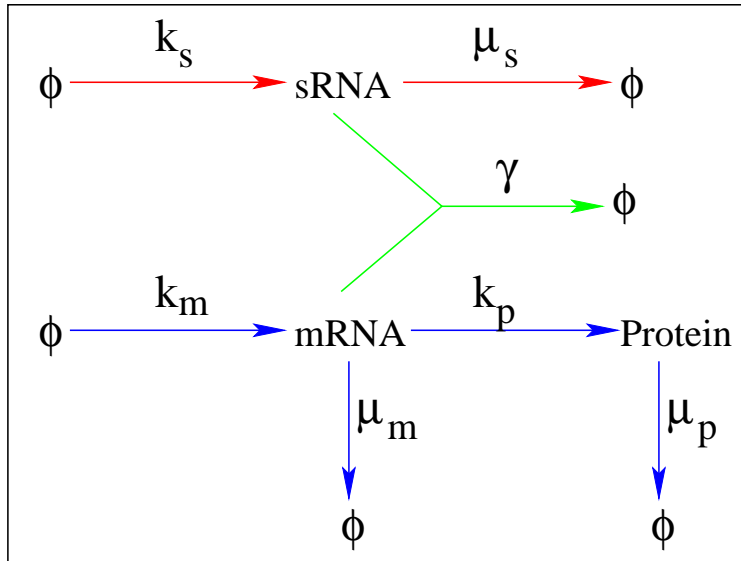


Figure 4. The sRNA-mRNA regulation scheme. The sRNA production rate is k_s and the degradation is μ_s . The interaction rate between the sRNA and mRNA that results in mutual degradation is γ . The mRNA and Protein reaction rates are the same as shown in figure 1.

The reaction scheme for small RNA based regulation has been studied by several groups [31, 32, 33, 34] and is schematically represented in Figure 4. In the limit of large concentrations of the small RNA regulator, the fluctuations of the small RNAs can be neglected and a more general model can be analyzed [20]. However, when the fluctuations of the regulator cannot be neglected, the exact solution of the model represented in Figure 4 is analytically intractable and approximations schemes need to be employed. In the following, we show how, in the limit of infrequent protein bursts, an analytical expression for the generating function of the protein burst distribution can be derived which agrees well with simulations.

We consider the case that mRNA production is governed by a Poisson process with constant rate k_m . In the limit of low k_m , the small RNA distribution *prior* to each burst can be well approximated by the unregulated small RNA distribution, which corresponds to a Poisson distribution with mean $n_s = \frac{k_s}{\mu_s}$. With these approximations, it is possible to derive an expression for the regulated protein burst distribution due to interaction with small RNAs as shown below.

Let us begin with the initial condition ($t = 0$) corresponding to the arrival of a mRNA. The protein burst distribution corresponds to the number of proteins produced from this single mRNA until the time it is degraded, either naturally or due to interaction with small RNAs. Our approach will focus on first deriving an expression for the survival probability of the mRNA at time t ($S(t)$). Let us define $P_1(n, t)$ as the probability that the mRNA exists at time t (i.e. it has not been degraded) and the number of sRNAs is n . Then, the mRNA survival probability is given by $S(t) = \sum_{n=0}^{\infty} P_1(n, t)$. Let us now define the operator \hat{H}_s which acts as follows

$$\hat{H}_s P(n) \equiv k_s [P(n-1) - P(n)] + \mu_s [(n+1)P(n+1) - nP(n)]. \quad (27)$$

In terms of this operator, we can write down the Master equation for $P_1(n, t)$ as follows

$$\partial_t P_1(n) = \hat{H}_s P_1(n) - \mu_m P_1(n) - \gamma n P_1(n), \quad (28)$$

The corresponding initial condition is taken as

$$P_1(n, t=0) = e^{-n_s} \frac{n_s^n}{n!}, \quad (29)$$

$$(30)$$

where $n_s = (k_s/\mu_s)$ (i.e., Poisson distribution of sRNAs at time $t = 0$) as discussed above.

In order to solve the Eq. (28) let us once again define a generating function

$$G_1(x, t) \equiv \sum_{n=0}^{\infty} x^n P_1(n, t), \quad (31)$$

which satisfies the partial differential equation

$$\partial_t G_1(x, t) = (\hat{H}_s - \mu_m - \gamma x \partial_x) G_1(x, t), \quad (32)$$

$$G_1(x, 0) = \exp(n_s(x-1)). \quad (33)$$

Here the differential operator \hat{H}_s can be easily derived from the equation Eq. (27), namely $\hat{H}_s = (x-1)(k_s - \mu_s \partial_x)$. The value of the generating function $G_1(x, t)$ at point $x = 1$ corresponds to $\sum_{n=0}^{\infty} P_1(n, t)$, i.e., the survival probability $S(t)$ of the mRNA molecule at time t . This survival probability can be obtained by solving Eq. (32) using the method of characteristics (Appendix). We obtain

$$S(\tau) = \exp[-\alpha(1 - e^{-\tau}) - \beta\tau], \quad (34)$$

where we have defined the following dimensionless parameters:

$$\tau = (\mu_s + \gamma)t, \quad (35)$$

$$\alpha = \left(n_s - \frac{k_s}{\mu_s + \gamma} \right) \frac{\gamma}{\mu_s + \gamma}, \quad (36)$$

$$\beta = \frac{\mu_m}{\mu_s + \gamma} + \frac{\gamma k_s}{(\mu_s + \gamma)^2}, \quad (37)$$

We can now proceed and calculate the generating function $G_b(x)$ of the protein burst distribution. Since protein production occurs at a constant rate k_p during the mRNA lifetime, the number of proteins produced by a surviving mRNA in time t is given by the Poisson distribution, with the corresponding generating function given by $e^{k_p(x-1)t}$. Since the difference $S(t) - S(t + \delta t)$ of survival probabilities is the probability that the mRNA degrades within the time interval $\{t, t + \delta t\}$, we obtain the burst generating function as

$$G_b(x) = - \int_0^\infty dt \partial_t S(t) e^{k_p(x-1)t}. \quad (38)$$

Rewriting the burst size distribution in terms of dimensionless parameters results in the following integral form

$$G_b(x) = 1 - k(1-x) \int_0^1 dz z^{k(1-x)+\beta-1} e^{\alpha(z-1)}, \quad (39)$$

where k is yet another dimensionless parameter

$$k \equiv \frac{k_p}{\mu_s + \gamma}. \quad (40)$$

The burst distribution with sRNA regulation, Eq. (39), has some interesting features. We note that Eq. (39) predicts that the burst distribution depends on three dimensionless parameters, α , β , and k (equations (36), (37), and (40)) and the steady-state distribution (see Eq. (21)) only adds a dependence on k_m/μ_p . Thus the modulation of any of the kinetic parameters shown in Figure 4 (for fixed k_m/μ_p) should result in the same steady-state distribution so long as the modifications occur in such a way that α , β , and k remain constant (and model assumptions/approximations are valid). As shown in Table 1, we can choose very different kinetic parameters that give rise to the same values for α , β and k and the prediction is that the burst and steady-state distributions for these different parameter choices should collapse onto a single curve. To test this scaling prediction, we carried out stochastic simulations based on the Gillespie algorithm [35] for a range of parameters such that α , β , and k remain constant. From the simulations, we recorded the resulting steady-state distributions and compared it to the analytic result (see figure 5). For the choice of parameters noted, we observed that the burst distribution is close to and can be well fitted by a geometric distribution. For a geometric distribution, the steady-state and burst generating functions are related by $G_s(x) = (G_b(x))^{\frac{k_b}{\mu_m}}$. We used this approximation to obtain the analytical form of the steady-state generating function and derived the steady-state protein distribution $P_s(n)$ using this. As can be seen in Fig. 4, the corresponding analytical results are in good agreement with results from simulations. The simulation results are also consistent with the scaling prediction since the curves with different parameter choices all collapse onto

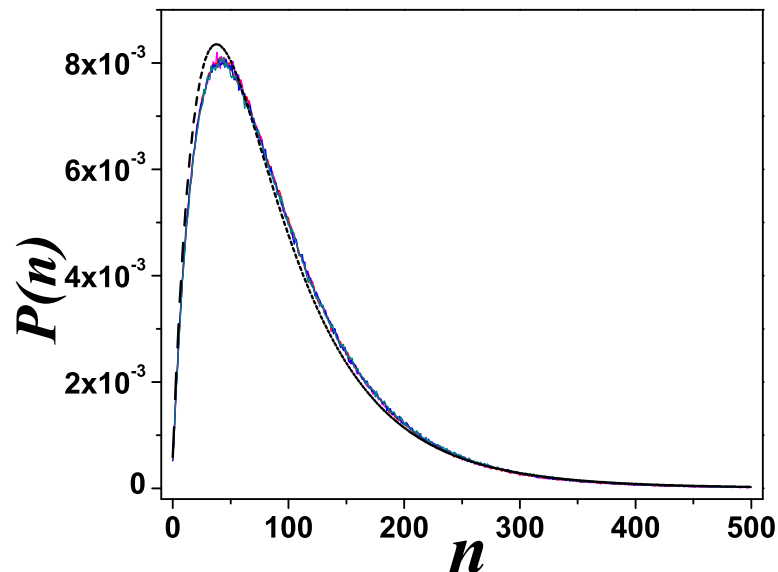


Figure 5. (color online) steady-state distributions with sRNA regulation. The dashed curve corresponds to Eq. (21) using Eq. (39) with approximations (see text) and $\alpha \simeq 4.76$, $\beta \simeq 1.34$, and $k \simeq 243.9$. The other curves are the results from four sets of numeric simulations. See Table 1 for the values of the parameters used in the simulations.

Simulation #	k_p	k_s	μ_s	γ
1	250	0.400313	0.072619	0.952381
2	300	0.717708	0.122308	1.107690
3	400	1.378120	0.217778	1.422222
4	500	2.055630	0.310870	1.739130

Table 1. The values of the parameters used in the numeric simulations shown in figure 5. For all simulations, $\alpha \simeq 4.76$, $\beta \simeq 1.34$, and $k \simeq 243.9$. Also, $\mu_m = 1$, $k_m = 0.01$, $\mu_p = 0.005$.

a single curve. The small discrepancy between the theoretical results and simulations is attributed to the approximations made, specifically the approximation for $G_s(x)$ noted above which is strictly valid only if the burst distribution is geometric. The results obtained from simulations of individual bursts are in very good agreement with the corresponding theoretical predictions.

7. Summary

Recent experiments underscore the need for connecting observed protein distributions from single-cell and single molecule studies using coarse-grained models of stochastic gene expression. In this context, several results have been derived in the present study which will help in the analysis of experimental results. We have shown how the functional form of the underlying mRNA burst distributions can be determined from observed protein distributions. If the protein burst distribution is geometric then the corresponding mRNA burst distribution has to be a conditional geometric distribution. The derived results further show that if the promoter can be repressed such that observed protein bursts arise from single mRNAs, then the underlying mRNA burst distribution in the unrepressed state can be completely determined. Furthermore, we derive relations connecting means and variances for burst and steady-state distributions for burst distributions which can deviate from a geometric distribution. The results derived also provide insight into regulation of protein expression bursts by small RNAs. The general results derived in this work can thus be used for analysis of a wide range of models of gene expression.

The authors acknowledge funding support from NSF (PHY-0957430) and from ICTAS, Virginia Tech.

8. Appendix

8.1. Relationship between mRNA and protein burst generating functions

Let us define $P(m, n; t)$ as the probability to find m mRNAs and n proteins after time t elapses since burst arrival. The corresponding generating function $G_p(x, y; t) \equiv \sum_{m,n} x^m y^n P(m, n; t)$ satisfies the following partial differential equation:

$$\partial_t G = \mu_m(1 - x)\partial_x G + k_p(y - 1)x\partial_y G. \quad (41)$$

The equation above can be easily solved by the method of characteristics

$$G(x, y; t) = G_m \left[\frac{1 - (1 + Yx)e^{-\mu_m Y t}}{Y} \right], \quad (42)$$

where $G_m[\cdot]$ is generating function of mRNAs at $t = 0$, and we defined

$$Y \equiv 1 - \frac{k_p}{\mu_m}(y - 1). \quad (43)$$

Therefore, the time dependent distribution of proteins in the burst is given by generating function

$$G_b(y; t) = G(1, y; t) = G_m \left[\frac{1 - (1 + Y)e^{-\mu_m Y t}}{Y} \right], \quad (44)$$

and the corresponding steady state is simply

$$G_b(y) = G_m \left[\frac{1}{Y} \right], \quad (45)$$

which is identical to the equation in the main text (Eq. (11)).

8.2. Two stage model

Assume that the mRNA production rate k_m has a value k_1 in the state 1 and a value k_2 in the state 2. The state 1 switches with probability λ_{12} into the state 2 and back with probability λ_{21} . One gets the following set of the equations for the generating functions $G_{1(2)}(x, t)$:

$$\partial_t G_1 = k_1 [G_b(x) - 1] G_1 + \mu_p(1 - x) \partial_x G_1 - \lambda_{12} G_1 + \lambda_{21} G_2, \quad (46)$$

$$\partial_t G_2 = k_2 [G_b(x) - 1] G_2 + \mu_p(1 - x) \partial_x G_2 + \lambda_{12} G_1 - \lambda_{21} G_2. \quad (47)$$

Let us explicitly calculate two moments of the protein's steady-state distribution. By setting $x = 1, t \rightarrow \infty$ we get

$$\lambda_{12} P_1^s = \lambda_{21} P_2^s, \quad (48)$$

$$P_1^s + P_2^s = 1, \quad (49)$$

where $P_1^s \equiv G_1(1, \infty)$ and $P_2^s \equiv G_2(1, \infty)$ are steady-state probabilities to be in the states 1 and 2 accordingly. Therefore, one derives

$$P_1^s = \frac{\lambda_{21}}{\lambda_{12} + \lambda_{21}}, \quad (50)$$

$$P_2^s = \frac{\lambda_{12}}{\lambda_{12} + \lambda_{21}}. \quad (51)$$

By evaluating the first derivative with respect to x at point $x = 1$ one can calculate $\langle n_i \rangle \equiv \sum_{n=0}^{\infty} n P_i(n)$, $i = 1, 2$:

$$0 = k_1 \langle n_b \rangle P_1^s - \mu_p \langle n_1 \rangle - \lambda_{12} \langle n_1 \rangle + \lambda_{21} \langle n_2 \rangle, \quad (52)$$

$$0 = k_2 \langle n_b \rangle P_2^s - \mu_p \langle n_2 \rangle + \lambda_{12} \langle n_1 \rangle - \lambda_{21} \langle n_2 \rangle. \quad (53)$$

Similarly, by evaluating the second order derivative with respect to x at point $x = 1$ one obtains

$$0 = k_1 [v_b P_1^s + 2 \langle n_b \rangle \langle n_1 \rangle] - 2 \mu_p \lambda_1 - \lambda_{12} v_1 + \lambda_{21} v_2, \quad (54)$$

$$0 = k_2 [v_b P_2^s + 2 \langle n_b \rangle \langle n_2 \rangle] - 2 \mu_p \lambda_2 + \lambda_{12} v_1 - \lambda_{21} v_2, \quad (55)$$

where we defined

$$v_i \equiv \sum_{n=0}^{\infty} n(n-1) P_i(n), \quad i = 1, 2 \quad (56)$$

$$v_b \equiv \sum_{n=0}^{\infty} n(n-1) P_b(n). \quad (57)$$

Hence, the average number of proteins in the steady-state and the variance can be derived by solving equations Eqs. (54,55) (result is given by the expressions Eqs. (24,25) in the main text.)

8.3. Derivation of survival probability for small RNA based regulation

Solution of the equation Eq. (32) using method of characteristics is given by

$$G_1(x, t) = \exp \left[-\beta\tau + \frac{k_s(x-1)}{\gamma + \mu_s} + \frac{k_s\gamma}{(\gamma + \mu_s)^2} \right] g(z), \quad (58)$$

where β and τ are dimensionless parameters as defined in the main text and the function $g(z)$ needs to be determined from the initial condition Eq. (33). Its argument is given by

$$z = \left[(x-1) + \frac{\gamma}{\gamma + \mu_s} \right] e^{-\tau}. \quad (59)$$

By matching the initial condition one gets

$$g(z) = \exp \left[-\frac{k_s}{\gamma + \mu_s} z \right] \exp \left[n_s z - \frac{\gamma n_s}{\gamma + \mu_s} \right]. \quad (60)$$

Finally, since we are interested in the quantity $S(t) \equiv G_1(1, t)$ (survival probability), we obtain

$$z \rightarrow \frac{\gamma}{\gamma + \mu_s} e^{-\tau}, \quad (61)$$

$$S(t) = \exp \left[-\beta\tau + \frac{k_s\gamma}{(\gamma + \mu_s)^2} \right] g(z). \quad (62)$$

from which the equation Eq. (34) from the main text can be obtained.

- [1] M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet*, 6(6):451–64, 2005.
- [2] J. M Paulsson. Models of stochastic gene expression. *Phys Of Life Rev*, 2(2):157–175, 2005.
- [3] Arjun Raj and Alexander van Oudenaarden. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell*, 135(2):216–226, 2008.
- [4] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden. Regulation of noise in the expression of a single gene. *Nat Genet*, 31(1):69–73, 2002.
- [5] M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci U S A*, 98(15):8614–9, 2001.
- [6] O. G. Berg. A model for the statistical fluctuations of protein numbers in a microbial population. *J Theor Biol*, 71(4):587–603, 1978.
- [7] J. Yu, J. Xiao, X. Ren, K. Lao, and X. S. Xie. Probing gene expression in live cells, one protein molecule at a time. *Science*, 311(5767):1600–3, 2006.
- [8] L. Cai, N. Friedman, and X. S. Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–62, 2006.
- [9] Y. Taniguchi, P. J. Choi, G. W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, 2010.
- [10] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36, 2005.
- [11] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi. Stochastic mrna synthesis in mammalian cells. *PLoS Biol*, 4(10):e309, 2006.
- [12] JR Chubb, T Trcek, SM Shenoy, and RH Singer. Transcriptional pulsing of a developmental gene. *Curr. Biol.*, 16(10):1018–1025, 2006.

- [13] B. B. Kaufmann and A. van Oudenaarden. Stochastic gene expression: from single molecules to the proteome. *Curr Opin Genet Dev*, 17(2):107–12, 2007.
- [14] Sandro Azaele, Jayanth R. Banavar, and Amos Maritan. Probing noise in gene expression and protein production. *Phys. Rev. E*, 80(3):031916, 2009.
- [15] V. Elgart, T. Jia, and R.V. Kulkarni. Application of Little’s Law to stochastic models of gene expression. *Phys. Rev. E.*, 82(2):021901, 2010.
- [16] Arjun Raj and Alexander van Oudenaarden. Single-Molecule Approaches to Stochastic Gene Expression. *Ann. Rev. Biophys.*, 38:255–270, 2009.
- [17] Daniel R. Larson, Robert H. Singer, and Daniel Zenklusen. A single molecule view of gene expression. *Trends in Cell Biology*, 19(11, Sp. Iss. SI):630–637, 2009.
- [18] N. Friedman, L. Cai, and X. S. Xie. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys Rev Lett*, 97(16):168302, 2006.
- [19] Piers J. Ingram, Michael P. H. Stumpf, and Jaroslav Stark. Nonidentifiability of the Source of Intrinsic Noise in Gene Expression from Single-Burst Data. *PLoS Comp Biol*, 4(10), 2008.
- [20] Tao Jia and Rahul Kulkarni. Post-transcriptional regulation of noise in protein distributions during gene expression. *Phys. Rev. Lett.*, 105(1):018101, 2010.
- [21] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, 3rd edition, 2007.
- [22] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. O’Shea, Y. Pilpel, and N. Barkai. Noise in protein expression scales with natural protein abundance. *Nat Genet*, 38(6):636–43, 2006.
- [23] John R. S. Newman, Sina Ghaemmamghami, Jan Ihmels, David K. Breslow, Matthew Noble, Joseph L. DeRisi, and Jonathan S. Weissman. Single-cell proteomic analysis of *S-cerevisiae* reveals the architecture of biological noise. *Nature*, 441(7095):840–846, 2006.
- [24] R Karmakar and I Bose. Graded and binary responses in stochastic gene expression. *Phys. Biol.*, 1(4):197–204, 2004.
- [25] Srividya Iyer-Biswas, F. Hayot, and C. Jayaprakash. Stochasticity of gene products from transcriptional pulsing. *Phys. Rev. E*, 79(3):031911, 2009.
- [26] Maciej Dobrzynski and Frank J. Bruggeman. Elongation dynamics shape bursty transcription and translation. *Proceedings of the National Academy of Sciences*, 106(8):2583–2588, 2009.
- [27] Vahid Shahrezaei and Peter S. Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, November 2008.
- [28] J. M. Pedraza and J. Paulsson. Effects of molecular memory and bursting on fluctuations in gene expression. *Science*, 319(5861):339–43, 2008.
- [29] Srividya Iyer Biswas. *Applications of Methods of Non-equilibrium Statistical Physics to Models of Stochastic Gene Expression*. PhD thesis, Ohio State University, 2009.
- [30] Lauren S. Waters and Gisela Storz. Regulatory RNAs in Bacteria. *Cell*, 136(4):615–628, February 2009.
- [31] E. Levine, Z. Zhang, T. Kuhlman, and T. Hwa. Quantitative characteristics of gene regulation by small rna. *PLoS Biol*, 5(9):e229, 2007.
- [32] Pankaj Mehta, Sidhartha Goyal, and Ned S. Wingreen. A quantitative comparison of sRNA-based and protein-based gene regulation. *Mol Sys Biol*, 4, 2008.
- [33] N. Mitarai, A. M. Andersson, S. Krishna, S. Semsey, and K. Sneppen. Efficient degradation and expression prioritization with small rnas. *Phys Biol*, 4(3):164–71, 2007.
- [34] V. Elgart, T. Jia, and R. V. Kulkarni. Quantifying mRNA synthesis and decay rates using small RNAs. *Biophys. J.*, 98(12):2780–2784, 2010.
- [35] D T Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361, 1977.